Title

DataUp: Further Development and Community Building

Permalink

https://escholarship.org/uc/item/1dt1v1pc

Authors

Cruse, Patricia
Strasser, Carly
Michener, William
et al.

Publication Date

2012

SUMMARY OF PROPOSED WORK

DataUp: Further Development and Community Building

Project Summary: Scientific data are increasingly born digital or are digitized early in the research process. Despite rapid growth in digital data, researchers rarely receive instruction in good data management practices. As a result, they often patch together idiosyncratic systems for data organization and documentation, which can have a negative impact on the long-term usability of their data. There is a rising demand for formalized data stewardship practices since funders and journals are now encouraging or mandating data management plans, data sharing, or both.

Recognizing that most Earth, environmental, and ecological scientists use spreadsheets at some point in the life cycles of their data, the California Digital Library (CDL) and its partners created a tool for Microsoft Excel that would encourage and enable good data stewardship practices. The result was a tool, DataUp, which facilitates documenting, managing, and archiving tabular scientific data.

Since its launch in fall 2012, response to DataUp has been enthusiastic. The CDL has received inquiries about DataUp from many repositories, organizations, and publishers interested in configuring the tool for their needs. These groups are most interested in customizing the DataUp tool for their user communities, which requires well-documented code, a familiar development platform, and an open API with adequate information for developers to use it. The CDL would like to see DataUp development move forward in order to capitalize on the potential opportunities for DataONE and beyond.

We are proposing a 12-month project to further develop the DataUp tool. The tool is useful as-is, but it has not reached its full potential as a tool for facilitating data management, sharing and archiving for researchers across disciplines. DataUp has the potential to become a key tool in research data sharing and archiving as envisioned by the NSF DataNet program. To that end, our major project goals for DataUp are to (1) enhance the tool?s user experience and add features, and (2) build the open-source community around DataUp. We plan to make improvements to DataUp via an iterative development process with community feedback and input. This community will include the existing DataUp user community, as well as researchers and information professionals from the University of California and DataONE.

Intellectual merit: DataUp will have a transforming effect on protecting the global scholarly community's investment in the "long tail" of research data. Much of this data is recorded in spreadsheets and produced in disciplines that have no organized approach to sharing and archiving, and have limited resource to do so (including citizen science groups). DataUp is the first tool that demonstrates such promise, and as the group that envisioned and built the original tool, the California Digital Library is uniquely qualified to fulfill that promise. Also, working with DataONE provides the best path forward to having a positive impact in the DataNet community.

Broader impacts: DataUp's repository- and discipline-agnostic design fosters an impact far beyond the Earth, environmental, and ecological sciences. Advancing data management and archiving practices in all disciplines will result in a more open scientific process, with readily available datasets that facilitate the progress of research. This has immeasurable benefits for society at large.

DataUp: Further Development and Community Building

1. Background

1.1 DataUp Project

Scientific datasets have immeasurable value, but they are rendered useless over time without proper documentation and long-term storage. Across disciplines as diverse as astronomy, demography, archeology, and ecology, large numbers of small heterogeneous datasets – the "long tail of data" [1] are especially at risk unless they are properly documented, saved, and shared. One unifying factor for many of these at-risk datasets is that they reside in spreadsheets.

In response to this need, the California Digital Library (CDL) partnered with Microsoft Research Connections and the Gordon and Betty Moore Foundation to create a data management software tool for Microsoft Excel. Many of the researchers creating these small, heterogeneous datasets use Excel at some point in their data collection and analysis workflow; we were interested in developing a data management tool that fits easily into these workflows and minimizes the learning curve for busy researchers.

The resulting DataUp project began in August 2011. We first formally assessed the needs of researchers by conducting surveys and interviews of our target research groups: earth, environmental, and ecological scientists. We found that, on average, researchers had very poor data management practices, were not aware of data centers or metadata standards, and did not understand the benefits of data management or sharing. Based on our survey results, we composed a list of desirable components and requirements and solicited feedback from the community to prioritize potential features of the DataUp tool. These requirements were then relayed to the software developers, and the DataUp tool was successfully launched in October 2012.

1.2 DataUp Tool

DataUp is free and open source, and has two forms: a web-based application (web app) and a downloadable Excel add-in. Both versions of the tool provide users with the ability to (1) perform a "best practices check" to ensure that data are well formatted and organized; (2) create standardized metadata (i.e., a scientifically-meaningful description of the data), using a wizard-style template; (3) retrieve a unique identifier for their dataset from their data repository; and (4) post datasets and associated metadata to a public data repository.

1. Best Practices Check. The tool determines whether the data file has any of the 11 potential issues that do not comply with data management best practices, such as embedded charts, comments, and color-coded cells. These issues were chosen based on interviews with researchers, as well as data managers who often "clean up" spreadsheets submitted by researchers for archiving. In addition to identifying the locations of these problems, DataUp explains why they are potentially problematic, and offers suggested alternatives or the ability to remove them in bulk.

2. Create metadata. DataUp will help the researcher create standard metadata using a form that becomes part of their spreadsheet, facilitating future use and sharing. Metadata can be generated at both the file- and column-level. File-level metadata includes names, email addresses and institutional affiliations for project personnel, and dataset titles. Column-level metadata (i.e. attribute metadata) includes information about the variables in the dataset, the units of measure, and descriptions of each column of data.

DataUp currently creates metadata using the Ecological Metadata Language (EML). This particular standard was chosen because of its widespread use in our original target communities. In addition, EML is both flexible and extensible, which enables future modifications to the chosen schema as necessary. We selected 47 elements of EML for DataUp, with only 6 required elements. We choose to support only a subset of EML in order to provide the lowest barrier to entry for researchers interested in documenting their datasets.

3. Obtain an identifier. Valuing and incentivizing the time and effort required to manage data well is an important factor in fostering data sharing and reuse. One way to allow data producers to get credit for this is through data citation. The DataUp tool connects to the chosen repository to retrieve a unique identifier for the researcher's dataset. Currently, this is done using the EZID service, based at CDL. The identifier generated is an ARK, which stands for "Archival Resource Key." ARKs provide stable, opaque, versatile, and transcription-safe identifiers. This identifier is saved in the data file's metadata.

4. Archive & share data. Once metadata is created, the user can connect directly to a repository via DataUp and upload their data for archiving. Currently, DataUp is connected to ONE*Share*, which is a dedicated public DataUp repository to which anyone may deposit tabular data (more information below).

1.3 DataUp Success

Response to the initial release of the tool has been enthusiastic. Within two months, the add-in version of the tool had been downloaded more than 280 times, and we estimate a proportionate interest in the web app version of the tool. The main DataUp website has had over 2,300 page views with visitors from more than 10 countries. These numbers do not, however, adequately represent the tool's popularity and potential. The CDL has received inquiries about DataUp from many repositories, organizations, and publishers interested in configuring the tool for their needs. . The inquiries represent a range of stakeholders that are crucial to data sharing, including a large citizen science project, a major social science data archive, some high-profile data publication services, and others. They are excited about the possibilities that DataUp represents for linking researchers' workflows directly to repositories, with capabilities for generating metadata and performing best practices checks.

*1.4 ONE*Share *Repository*

The current version of DataUp is configured for depositing data into the ONE*Share* repository, a publicly accessible repository created specifically for DataUp. ONE*Share* is a special instance of CDL's Merritt Repository, which serves as a digital archive to the 10 University of California campuses. The ONE*Share* repository has an added benefit of connecting to the NSF-funded

DataONE network of data centers [2]. DataONE links together existing data centers and enables its users to search for data across all participating repositories using a single search interface. Data deposited into ONE*Share* via the DataUp tool are indexed and made discoverable by any DataONE user, facilitating collaboration and enabling data re-use.

1.5 DataUp Architecture

Initial development of DataUp was funded by Microsoft Research Connections. DataUp's codebase is written in C# using the .NET application framework. The web app is deployed on Windows Azure (Microsoft's cloud platform).

DataUp's architecture (Figure 1) consists of two clients communicating via a web service to one or more repositories. The add-in client is an Excel extension that runs directly on a researcher's Windows-based computer. The web app client (http://dataup.org/) runs as an online application hosted in Microsoft Azure (http://windows.azure.com/). Both clients support the Ecological Metadata Language (EML, http://knb.ecoinformatics.org/software/eml/) and draw functionality from a common web service, also hosted in Azure. That web service is managed by a separate administrative service.

DataUp was designed not only for standalone metadata checks, but also for contacting a variety of repositories to obtain persistent identifiers and to archive data. Currently, the only repository supported is ONEShare, an instance of a DataONE Member Node that uses the Merritt API. With the front-end running at CDL and a storage node back-end running at the University of New Mexico, content can be browsed by logging in to Merritt as a guest (http://merritt.cdlib.org/) and discovered via DataONE (https://cn.dataone.org/onemercury).

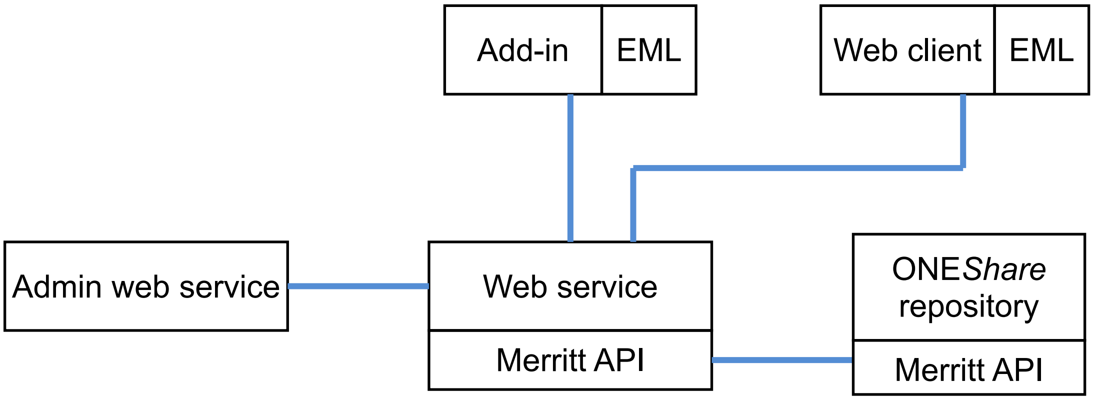

Figure 1. The current DataUp architecture

2. Project Intent & Goals

DataUp is now operational, but it has not reached its full potential as a tool for archiving datasets from a range researchers and disciplines. DataUp has the potential to become a key tool in research data sharing and archiving as envisioned by the NSF DataNet program. To that end, our major project goals for DataUp are to (1) enhance the tool's user experience and add features, and (2) build the open-source community around DataUp.

We plan to make improvements to DataUp via an iterative development process with community feedback and input. The CDL's primary project responsibility for the original DataUp project was to interact with the user community and build software requirements; we are therefore familiar with the community, and a level of trust has been established that we will build upon in the course of this project. The goals of this project would result in our ability to foster this open source community.

2.1 Deliverables

Project outcomes include a set of baseline features and enhancements that are clearly useful to DataUp based on the collective experience of its use to date, and a further set of features to be proposed and evaluated in consultation with the DataUp user community during the project itself. The major project deliverables are:

- Development of an engaged and active user and open source development community. This will have a critical impact on the long-term adoption and sustainability of the DataUp tool.
- Support for alternative user identity providers, including Google, Twitter, and InCommon. This will provide DataUp users with greater flexibility in authenticating themselves to the web application using already existing institutional and social media credentials. Providing support for additional identity providers, for example, CILogin, may be pursued if determined to be a priority by the user community.
- Changes to the language used by the DataUp add-in and web application user interfaces to remove potential ambiguity and enhance user experience based on feedback received from use to date. There will be continual evaluation of new development prototypes throughout the course of the project to ensure the highest level of user experience.
- Support for alternative repositories, including DataONE member nodes using the native DataONE member node API. This will provide DataUp users with greater flexibility in selecting institutional or disciplinary service providers for stewarding their data. Providing support for additional repositories, for example, those compliant with the common SWORD API, may be pursued if determined to be a priority of the user community.
- Validation of user-provided email address prior to posting spreadsheets to a repository. This will protect DataUp users from missing important notifications dues to improperly transcribed addresses and repository managers from potential spamming attacks.
- Improvements to metadata handling, including:
    - Support for a fuller subset of the EML schema. This will afford DataUp users with the ability to provide richer scientific description of their data.

  - Enforcement of conditional metadata requirements, that is, ensuring the presence of certain metadata elements that are required only in concert with other optional elements. This will ensure the consistent generation of schema-validated metadata.

  Support for metadata schemas beyond EML may be pursued if determined to be a priority of the user community.

- Improvements in the best practices check. The specifics of this task will be defined through consultation with the DataUp user community, and in particular, with repository managers and preservation analysts, who will be most familiar with spreadsheet features that may be inimical to the long-term stewardship and sharing of tabular data.

Beyond the specific deliverables described above, there are two further work items that will be investigated during the project:

- Investigation the development of an alternative DataUp add-in suitable for the Mac versions of Excel. As a significant proportion of the researcher community relies on Mac, rather than Windows, computers, this would have a strong impact on the adoption of the DataUp tool. However, it must be noted that the development support for Mac Excel add-ins is not nearly as comprehensive or robust as for Windows.
- Investigation of the reimplementation and rehosting of the DataUp add-in, web application, and mediating web service in more mainstream alternatives to C#, .NET, and Azure. If moved forward, this can have a significant impact on the long-term sustainability of DataUp, as it would become much more accessible to a larger community of open source developers.

As noted above, a certain set of project activities will be directed by consensus agreement of the project team and the DataUp user community. As a finite amount of project time will be available for this work, it is possible that some of the items given lower priority may not be fully implemented during the project duration.

3 Project Structure

We propose a project duration of 12 months, with a focus on development informed by the DataUp user community including the University of California and DataONE, resulting in an improved tool with a widespread user base and a engaged open-source developer community.

3.1 Development

The proposed work will introduce flexibility at a number of different points in the architecture shown in Figure 1, such as adding new repository types, new metadata types, and perhaps new hosting (e.g., AWS) and client (e.g., Mac) platforms. Development will occur in four phases, each informed by the community's input and feedback. Development priorities for the first phase will focus on feedback from beta testers of the current DataUp tool [3,4]. The tasks are arranged in tiers according to estimated effort, and they include improvements, bug fixes, and stability issues. Beyond Tier 1, the task list should be considered provisional and subject to change based on the assessment component.

Tiered Development Phases	
Tier 1	• Develop regression tests and agile work practices. • Develop deployment procedures for a dedicated development and testing instance of DataUp.
Tier 2	• Add alternative user identity providers to the web application. • Make small changes to user interface (UI) such as wording, labels, and language • Add a repository type for repositories that conform to a DataONE API.
Tier 3	• Validate email provided by the user before allowing deposit (this will protect repositories from spam and protect users from typos that would result in lost archive responses). • Add additional repository types. • Expand EML to support richer description of the tabular data (additional elements and more sophisticated user prompts). • Modify the UI to enforce conditional requirements. • Improve metadata generation. • Improve the tool's best practices check.
Tier 4	• Add additional metadata schemas. • Investigate building an add-in for Mac operating systems. • Investigate the potential of conversion of the code from its C#-.NET-Azure base to more mainstream alternatives more likely to attract developers from the open source community. • Allow for expansion and customization of the best practices check.

3.2 Community involvement

All of the proposed DataUp development cycles will be driven by community feedback. In addition to the initial list of features and improvements resulting from the original DataUp project, we will have two user events at strategic points in development to collect input on the tool's progress.

The DataUp project has had close ties to the researcher community since its inception: requirements were built based on interviews and surveys, a blog and Twitter feed provide project updates and opportunities for community suggestions, and the resulting code is open source and freely available. We plan to continue fostering this climate of community input throughout the proposed project using these channels:

- Updates to the DataUp website and documentation on code site [5]
- Presentations and demonstrations at conferences and invitational meetings
- Blog updates

The user events will provide direction and focus for future development priorities. The first event will be in the form of a workshop or focus group: we will assemble researchers from DataONE and the University of California, in varying stages of their careers and from a range of disciplines. These potential users will be asked to interact with the tool and provide detailed information about usability and features. The second user event will be a virtual meeting, wherein we will first describe how participants should test DataUp, followed by a period of testing and feedback. The virtual meeting will include those who participated in the focus group, as well as anyone else interested in participating.

4. Personnel roles and qualifications

The project has institutional support from the CDL. The CDL supports the work of Patricia Cruse, Director, University of California Curation Center, California Digital Library on the project and she will lend her expertise to the project.

The CDL also supports John Kunze's, University of California Curation Center, California Digital Library work on the project. He will provide technical oversight for the project. He served in this role for the original DataUp project and is therefore well-placed to help guide the technical aspects of the development process. Kunze designed the ARK identifier scheme, created the DOI and ARK resolver infrastructure behind the EZID system, was a principal author of the Dublin Core metadata standard, and is on the leadership and core cyberinfrastructure teams for DataONE.

The project calls for one full-time developer for a period of 12 months (to be hired). This person will be responsible for implementing all code changes and modifications, and any other technical aspects of the DataUp development. This individual will be based at CDL.

The Project manager (to be hired) will be responsible for setting the development priorities, organizing and implementing the focus groups, communicating with the community, and promoting the DataUp tool. We expect this to be a part-time position based at CDL.

Patricia Cruse, PI and Director, University of California Curation Center, CDL will guide the project. Other personnel include William Michener, Rebecca Koskela, David Vieglais, and Amber Budden of the DataONE project. These individuals will help guide both development and oversight via frequent communication with the personnel above. The CDL would work collaboratively with the DataONE developer team to ensure compatibility with DataONE infrastructure.

5. Timeline

Month	Tasks
1	Prioritize development; community feedback Familiarize with code
1-5	Development Phase 1
5	Focus group and testing on Development Phase 1 product Revise development prioritization
6-8	Development Phase 2 Focus group and beta testing on Development Phase 2 product Revise development prioritization
9-10	Development Phase 3 Focus group and beta testing on Development Phase 3 product

Month	Tasks
	Revise development prioritization
11-12	Final Development Phase Integration, documentation DataUp Version 2 promotion and release

6. References

[1] Heidorn, PB. 2008. Shedding Light on the Dark Data in the Long Tail of Science. Library Trends 57 (2): 280-299

[2] Michener, W et al. DataNetONE (Observation Network for Earth). NSF Grant No. OCI 0830944.

[3] DataUp Wiki site of issues and improvements based on beta testers: https://bitbucket.org/dataup/main/wiki/improvements_issues

[4] Open Issues for DataUp development: https://bitbucket.org/dataup/main/issues?status=new&status=open

[5] DataUp BitBucket site: https://bitbucket.org/dataup/main